Università degli Studi di Padova

Department of Industrial Engineering

Machine learning
Lesson #4

Academic year 2025-2026

Prof. Pierantonio Facco

CAPE-Lab, Computer-Aided Process Engineering Laboratory

Email: pierantonio.facco@unipd.it

URL: https://research.dii.unipd.it/capelab/

Important notice on PLS_Toolbox® and Minitab®

- A license of the **Minitab®** is available

- A demo license of the last version of the **PLS_Toolbox®** is available
  - it expires after 6 months
  - 6 extra months will be available when necessary
  - **kindly delivered by Eigenvector Research Inc. for free**

- **The software could be used only to the purpose of completing the exam of MLfPE!**
  - the software is **strictly confidential**
  - any other use is strictly forbidden
  - the disclosure of the software will be legally pursued
  - use your UniPD email to assess the software

Recap of the previous lesson

- We are dealing with the simple case of **univariate data**
- We learned what are **random variables**, PDFs and what are effective statistical indices to describe it
- We introduced a special case (and very common one) of random variable, the **Gaussian** one:
  - the concepts of **PDF, CDF and inverse density function** were introduced to relate the occurrence of an event to the probability of occurrence

Today's lesson: important distributions

- Not only Gaussian distributions are important
- Other common distributions are widely employed in practice (e.g.: in SPC, DoE, hypothesis testing, etc.):
  - Students' **t distribution**
  - Chi-square χ^2 **distribution**
  - Fisher's **F distribution**
  - all of them will be very useful for the entire course!
    - and also for your professional life… :)
- **Hypothesis testing** will be studied

χ^2 distribution

χ^2 distribution (chi^2 distribution)

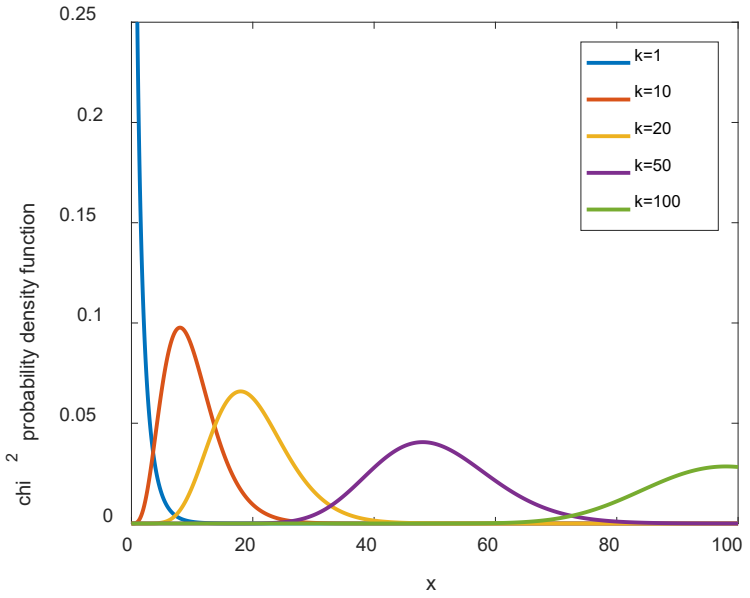
- The **χ^2 distribution** is an important probability distribution defined in terms of **normal random variables**:
  - if $z_1, z_2, \ldots, z_K$ are K normally and independently distributed random variables with mean 0 and variance 1, abbreviated: $z_k \sim NID(0,1)$ for all the $k = 1, 2, \ldots, K$,

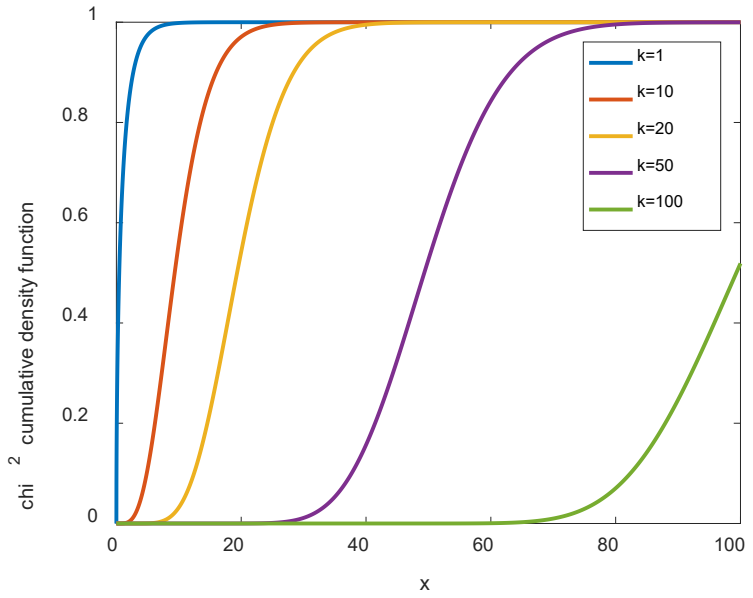
    then

  the random variable:

$$x = z_1^2 + z_2^2 + \cdots + z_K^2$$

  is a χ^2-distribution with K degrees of freedom





PDF of a χ^2 distribution

- The **probability density function** of the χ^2 distribution is:

$$f(x) = \frac{1}{2^{K/2}\Gamma\left(\frac{K}{2}\right)} x^{\frac{K}{2}-1} e^{x/2}$$

where $\begin{cases} \Gamma\left(\frac{K}{2}\right) = \sqrt{\pi}\frac{(k-2)!}{2^{\frac{k-1}{2}}} & \text{for even } k \\ \Gamma\left(\frac{K}{2}\right) = (k/2-2)! & \text{for odd } k \end{cases}$

  - the mean and the variance are, respectively:

$$\mu = K$$
$$\sigma^2 = 2K$$

  - the shape of the distribution is skewed
  - the Matlab® commands for:
    - the χ^2 PDF is: **`chi2pdf`**
    - the χ^2 CDF is: **`chi2cdf`**
    - the inverse of the χ^2 DF is: **`chi2inv`**
    - χ^2 distributed random numbers: **`chi2rdn`**

Typical example of χ^2 distribution

- Suppose that $w_1, w_2, \ldots, w_N$ is a random sample from a $N(\mu, \sigma^2)$ distribution, then:

$$\frac{SS}{\sigma^2} = \frac{\sum_{i=1}^{N}(w_i - \bar{w})^2}{\sigma^2} \sim \chi^2_{N-1}$$

  - SS/σ^2 is distributed as χ^2 distribution with $(N-1)$ degrees of freedom
- This is extremely important and occurs repeatedly, because *a lot of statistical methodologies (e.g.: statistical process control, design of experiments, etc…) involve the computation/manipulation of sums of squares*
- For example, the sample variance can be written as:

$$s^2 = \frac{SS}{N-1}$$

  - if the observations in the sample are $NID(\mu, \sigma^2)$ then the distribution of s^2 is:

$$\frac{\sigma^2}{N-1}\chi^2_{N-1}$$

  - thus, if the population is normally distributed, the sampling distribution of the sample variance is a constant times the chi-square distribution
  - verify it in Matlab®

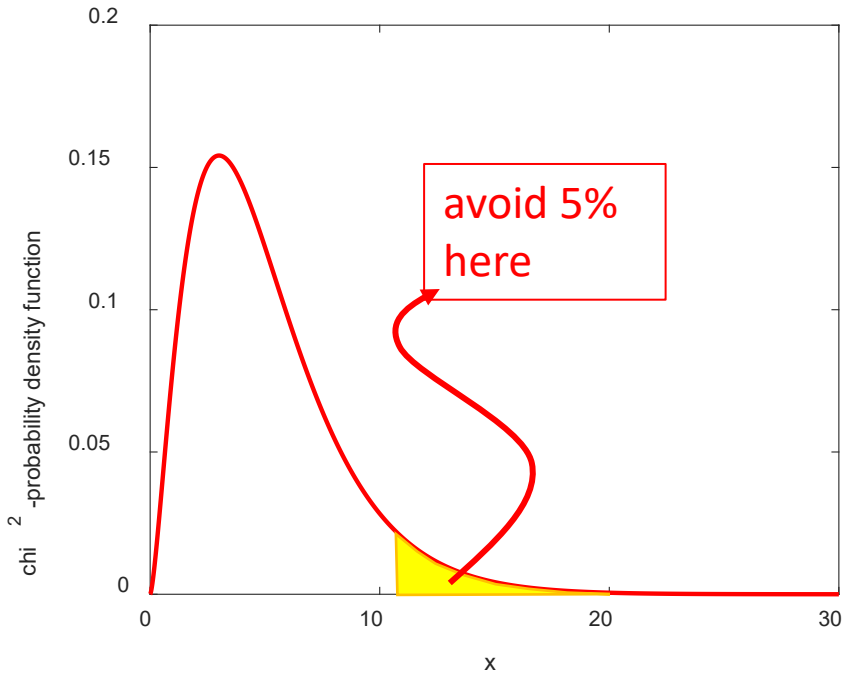
Percentage points of the χ^2 distribution

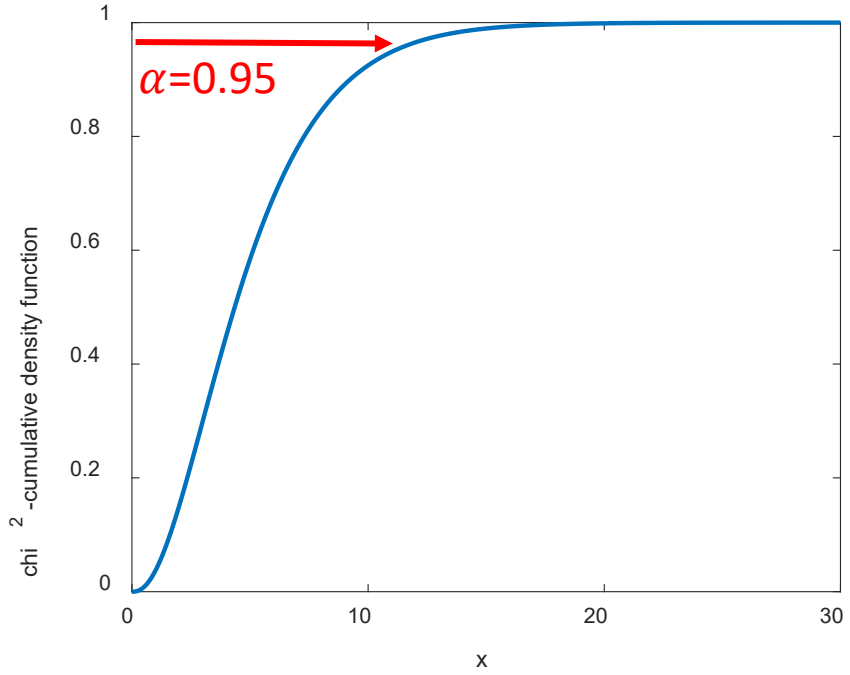
- Tables on the percentage points of the χ^2 distribution can be found in the textbooks, depending on:
  - the number of degrees of freedom K (rows of the table)
  - the selected percentile α (columns of the table)

K \ α	0.995	0.990	0.975	0.950	0.500	0.050	0.025	0.010	0.005
1	0.00 +	0.00 +	0.00 +	0.00 +	0.45	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	1.39	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	2.37	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	3.36	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	4.35	11.07	12.38	15.09	16.75
6	0.68	0.87	1.24	1.64	5.35	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	6.35	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	7.34	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	8.34	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	9.34	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	10.34	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	11.34	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	12.34	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	13.34	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	14.34	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	15.34	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	16.34	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	17.34	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	18.34	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	19.34	31.41	34.17	37.57	40.00

Example: inverse χ^2 distribution (1/2)

- Where is located the limiting value of a χ^2 distributed random variable with 5 d.o.f. that guarantees that the $100(1-\alpha) = 95\%$ percent of the lowest observations are within that limiting value?
  - `chi2inv(0.95,5)` ➡ 11.0705





Example: inverse χ^2 distribution (2/2)

- Where is located the limiting value of a χ^2 distributed random variable with 5 d.o.f. that guarantees that the 95% percent of the lowest observations are within that limiting value?

K \ α	0.995	0.990	0.975	0.950	0.500	0.050	0.025	0.010	0.005
1	0.00 +	0.00 +	0.00 +	0.00 +	0.45	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	1.39	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	2.37	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	3.36	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	4.35	11.07	12.38	15.09	16.75
6	0.68	0.87	1.24	1.64	5.35	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	6.35	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	7.34	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	8.34	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	9.34	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	10.34	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	11.34	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	12.34	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	13.34	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	14.34	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	15.34	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	16.34	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	17.34	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	18.34	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	19.34	31.41	34.17	37.57	40.00

t distribution

t-distribution

- If z is a standard normal variable, χ^2_K is a χ^2 distributed random variable with K degrees of freedom and they are independent, then:

$$t_K = \frac{z}{\sqrt{\frac{\chi^2_K}{K}}}$$

follows the t **distribution with** K **degrees of freedom** t_K.

- The **PDF of the t distribution** is:

$$\mathrm{f}(x) = \frac{\Gamma\left(\frac{K+1}{2}\right)}{\sqrt{K\pi}\,\Gamma\left(\frac{K}{2}\right)}\left(\frac{x^2}{K}-1\right)^{-\frac{K+1}{2}} \quad \text{for} \quad -\infty < x < +\infty$$

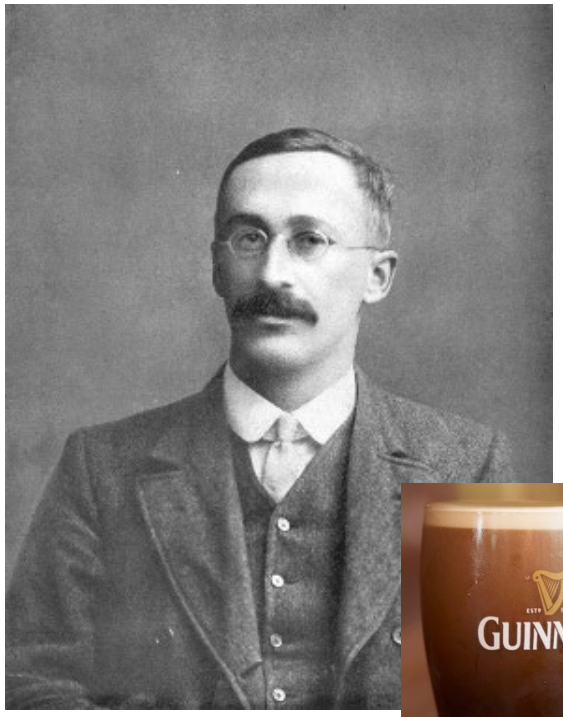
- The **mean** and **variance** are:

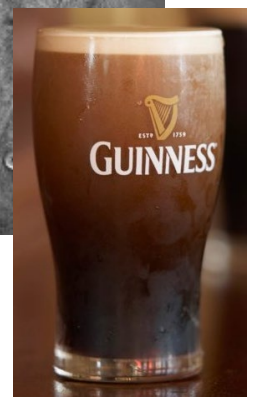
$$\mu = 0$$

$$\sigma^2 = \frac{K}{K-2} \quad \text{for} \quad K > 2$$

Student’s t distribution

- **William Sealy Gosset** (13 June 1876 – 16 October 1937) was an English statistician
- He developed the t-distribution
- The t-distribution is known to be the **Student’s t distribution**

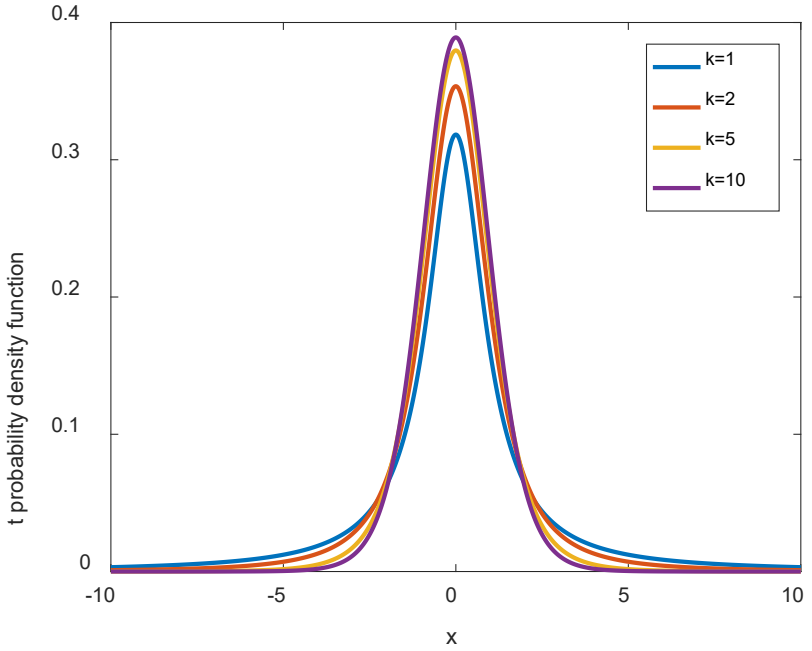


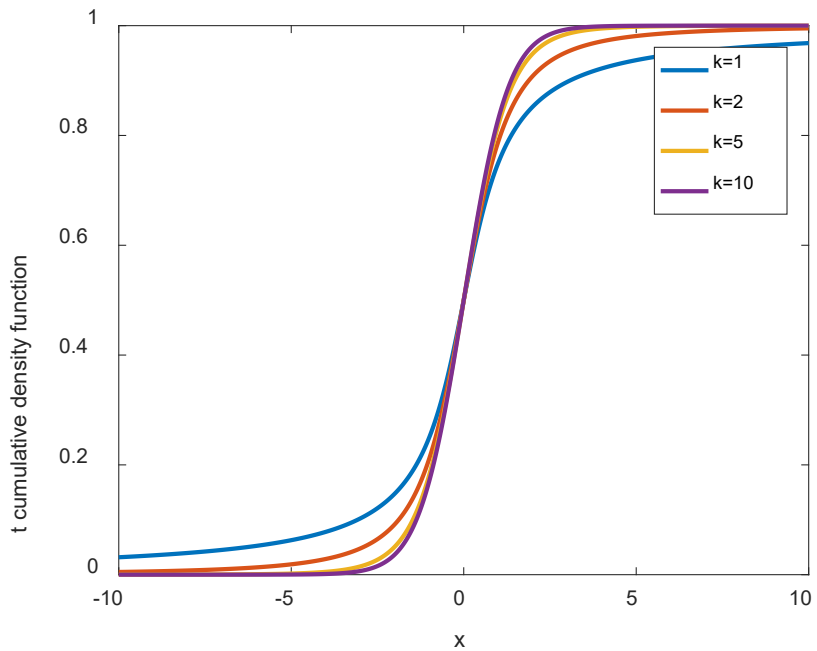


- He studied Chemistry and Mathematics at the New College in Oxford
- In 1899 he joined the **Arthur Guinness & Son brewery** in Dublin
  - as an employee of Guinness, Gosset applied his statistical knowledge both in the brewery and on the farm, as you could do after the course of Machine learning for process engineering ;)
- Another researcher at Guinness had previously published a paper containing trade secrets of the Guinness brewery
  - to prevent further disclosure of confidential information, Guinness prohibited its employees from publishing any papers
- After pleading with the brewery and explaining that his mathematical and philosophical conclusions were of no possible practical use to competing brewers, he was allowed to publish them, but under a pseudonym the "Student", to avoid difficulties with the rest of the staff :D

PDF and CDF of the t distribution

- The PDF of the t distribution is a bell-shaped distribution, always centered in $x = 0$
  - the lower the parameter K, the flatter the shape of the PDF
  - it is a sort of heavy-tailed standard normal distribution
- The CDF is a sigmoid line that also depends on the parameter K





Examples of t-distributions

- If that $w_1, w_2, \ldots, w_N$ is a random sample from $N(\mu, \sigma^2)$ then:

$$t = \frac{\bar{w} - \mu}{\frac{s}{\sqrt{N}}}$$

is distributed as a t distribution with $K = (N - 1)$ degrees of freedom

  - verify it in Matlab®

- If the parameter $K \to \infty$, the *t* distribution becomes **the standard normal distribution**:
  - for $K > 50$ the t distribution is not even distinguishable from the standard normal one
    - verify it in Matlab®
- As in the case of the other distributions the **percentage points of the *t* distribution** are given in tables that can be commonly found in the statistics' textbooks

$K \backslash \alpha$	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.727	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.019	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587

Everyday homework

- Compare in Matlab® the shape of a t distribution to the standard normal distribution:
  - use commands:
    - `tpdf`
    - `tcdf`
- Use the inverse density function to understand where the 95% confidence limits are located using command: `tinv`
  - compare the result with the one you can obtain with the percentage point tables, for example in the case when $K = 5$
    - `tinv(0.975,5)` ➡ 2.5706
    - `tinv(0.025,5)` ➡ -2.5706

K \ α	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.727	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.019	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587

F distribution

F distribution

- If χ_N^2 and χ_M^2 are two independent random variables of χ^2 distributions with N and M degrees of freedom, respectively, then:

$$F = \frac{\frac{\chi_N^2}{N}}{\frac{\chi_M^2}{M}}$$

follows the F **distribution with N numerator degrees of freedom and M denominator degrees of freedom**

- The **probability distribution** of a random variable x that is F distributed is:

$$f(x) = \frac{\Gamma\left(\frac{N+M}{2}\right)\left(\frac{N}{M}\right)^{\frac{N}{2}} x^{\frac{N}{2}-1}}{\Gamma\left(\frac{N}{x}\right)\Gamma\left(\frac{M}{2}\right)\left[\frac{N}{M}x+1\right]^{\frac{N+M}{2}}} \quad \text{for} \quad 0 < x < +\infty$$

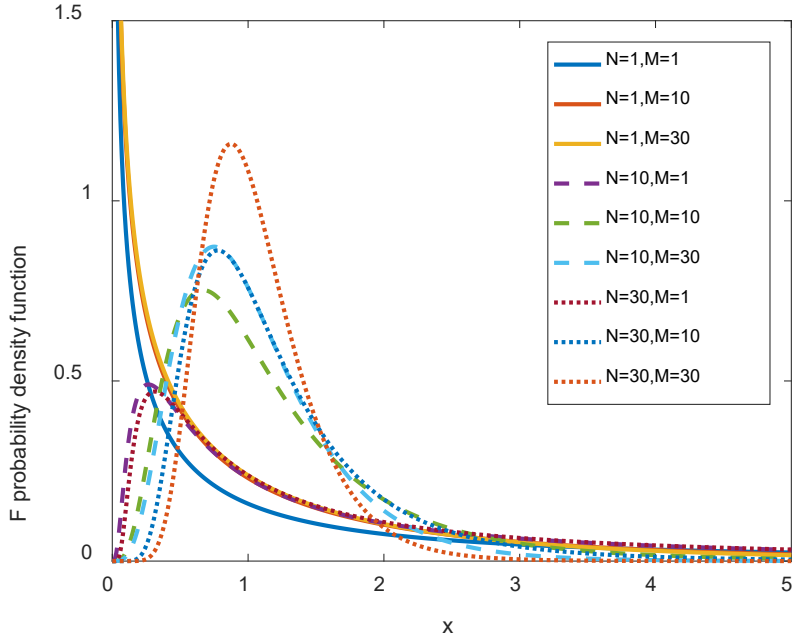
- The **mean** and **variance** are, respectively:

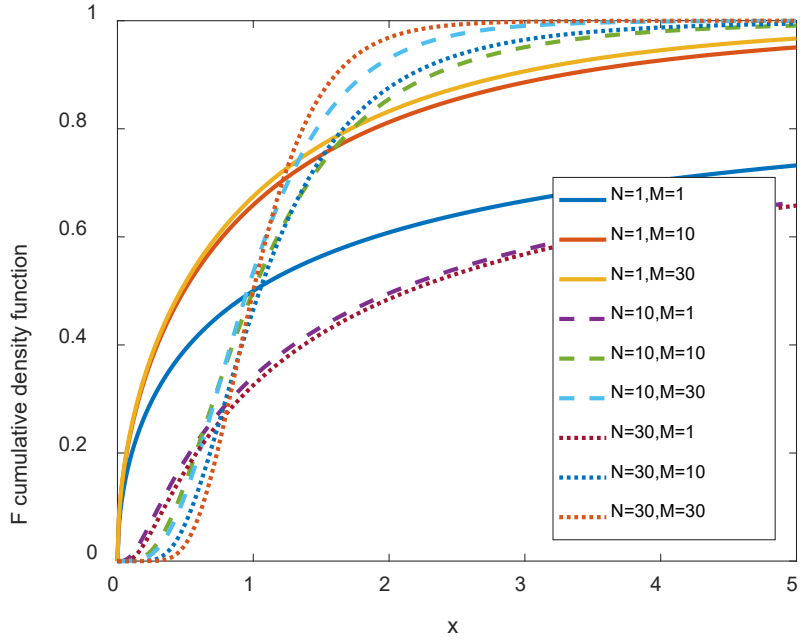
$$\mu = \frac{M}{M-2} \quad \text{if } M > 2$$

$$\sigma^2 = \frac{2M^2(N+M-2)}{N(M-2)^2\,(M-4)} \quad \text{if } M > 4$$

Fisher's F distribution and its PDF and CDF

- This distribution is very important in the statistical **analysis of designed experiments** and **statistical process control**:
  - the distribution was presented by the father of the Design of Experiments (DoE)
- Examples of the PDF and the CDF are reported for different values of the parameters N and M





Examples of F distribution

- Example: suppose to have two independent normal populations with common variance σ^2:
  1. $w_{11}, w_{12}, \ldots, w_{1N}$ is a random sample of N observations from the population #1
  2. $w_{21}, w_{22}, \ldots, w_{2M}$ is a random sample of M observations from the population #2

then

$$\frac{s_1^2}{s_2^2} \sim F_{N-1,M-1}$$

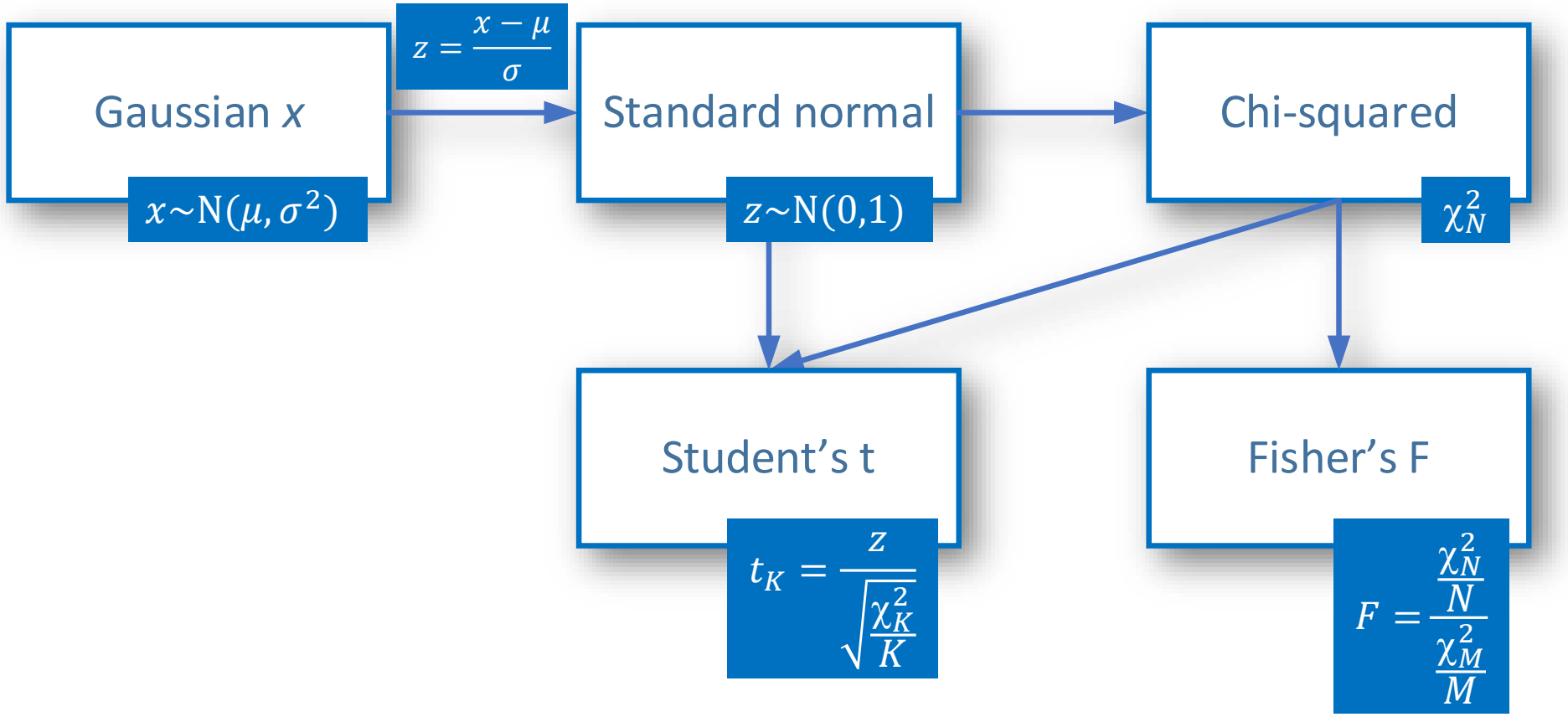
the **ratio between the sample variances** of the two populations s_1^2 and s_2^2 follows a F distribution with $N-1$ and $M-1$ degrees of freedom

- **Percentage points** of the *F* distribution are given in tables which are usually available in the textbooks
- In Matlab® try to practice with commands:
  - `fpdf`
  - `fcdf`
  - `finv`

$\alpha = 0.1$

M \ N	Degrees of Freedom for the Numerator N																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47

Summary of the relation between random variables



Hypothesis testing

Statistical hypothesis

- A **statistical hypothesis** is a statement either about the parameters of a probability distribution or the parameters of a model:
  - the hypothesis reflects some **conjecture** about the problem situation

- The hypothesis testing is stated formally as:

$$H_0: statement\ of\ null\ hypothesis$$

$$H_1: statement\ of\ alternative\ hypothesis$$

  - H_0 is called the **null hypothesis**
  - H_1 is called the **alternative hypothesis**
    - this can be either a **one-sided** or a **two-sided alternative hypothesis**

Procedure for hypothesis testing

- **Hypothesis testing**:
  1. taking a **random sample**
  2. computing an appropriate **test statistic**
  3. identifying the **critical region** (i.e., **rejection region**) for the tests with a predetermined significance
     - specifying the set of values for the test statistic that leads to rejection of H_0 through the confidence limits
  4. **rejecting or failing to reject the null hypothesis** H_0 based on the computed value of the test statistic

- Two kinds of errors may be committed when testing hypothesis:
  - **type I error**:
    - the null hypothesis is rejected when it is true
  - **type II error**:
    - the null hypothesis is *not* rejected when it is false
  - special symbols are given to the probabilities of these two errors

$$\alpha = \mathrm{P}(\text{type I error})$$
$$\beta = \mathrm{P}(\text{type II error})$$

    - sometimes it is more convenient to work with the **power** of the test $(1-\beta)$
- The **significance level** of the test, namely the probability of type I error, should be found
  - design the test so that the probability of type II error has a small value

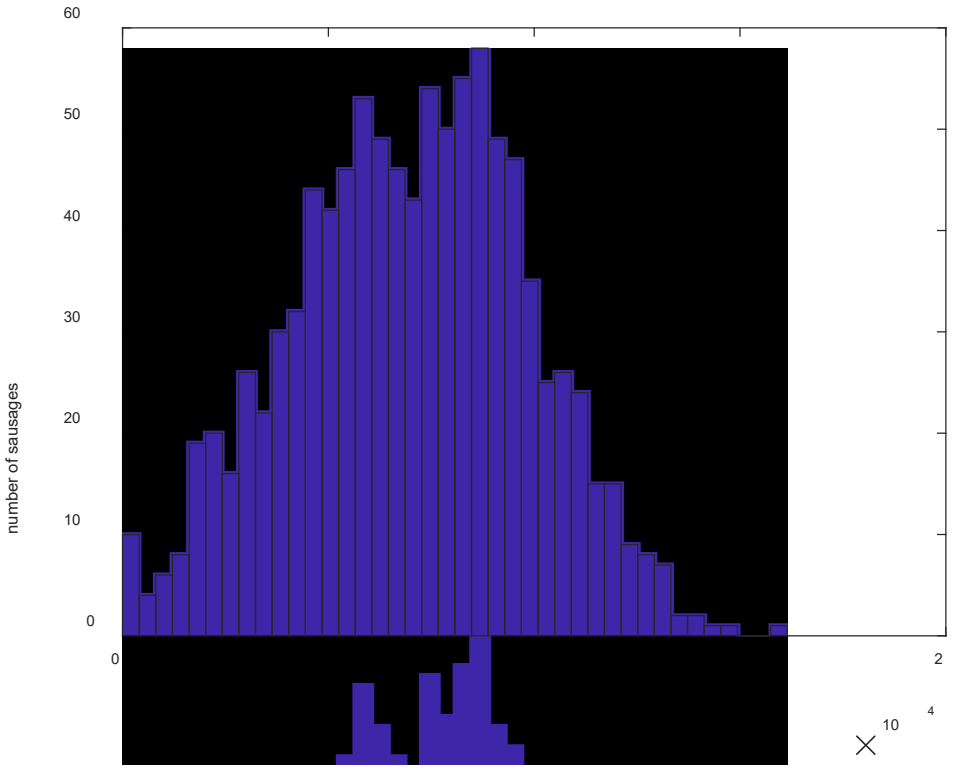
Industrial applications of probability theory and hypothesis testing

Example of probability theory: food industry

- Example:
  - you work for a food industry producing sausages
  - an extended experimental campaign was carried out to evaluate the contamination of sausages
    - the outcome is summarized in the data file: **sausage.xlsx**
  - your manager wants to know
    1. what is the natural variability of the staphylococcus contamination of the sausages?
    2. what is the percentage of the product which is out of the specification limits of 10 000 c.f.u./g of staphylococcus?

- Solution:
  1. the average contamination is: 6992.5c.f.u./g with a standard deviation of: 2865 c.f.u./g
  2. the probability of finding a sausage which is out of specification (i.e., has a contamination of >10 000 c.f.u./g od staphylococcus) is:

$$P(x > 10000\ c.f.u./g\ ) = 0.1469$$



Example of hypothesis testing on Student t: water contamination (1/2)

- Example:
  - you work in the field of water purification
  - you want to understand if the bacteriological contamination of water in terms of $\log_{10}$(c.f.u./mL) can be considered statistically equal to 5, as prescribed by the low, with a significance $\alpha = 0.01$
    - the outcome is summarized in the data file: **water_contamination.xlsx**

- Solution:
  - the hypothesis to be tested is:

$$\left.\begin{array}{l} \mathrm{H}_0\colon\ \mu = 5 \\ \mathrm{H}_1\colon\ \mu \neq 5 \end{array}\right\}$$

2-sided test because we want to exclude that both $\mu > 5$ and $\mu < 5$

  - a t-test should be performed
    - the observed value of t is:

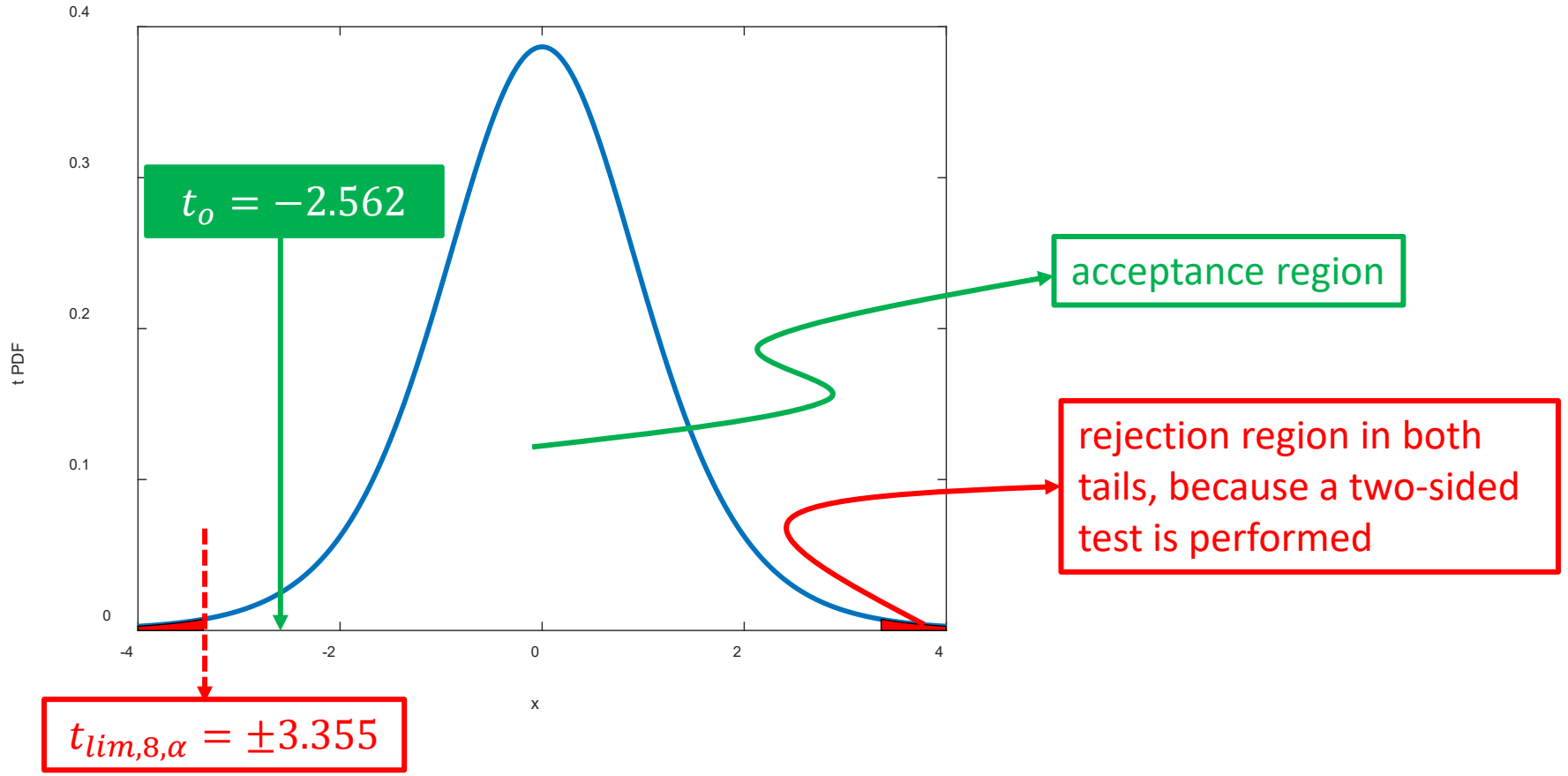
$$t_o = \frac{\bar{x} - \mu}{s/\sqrt{N}} = \frac{4.789 - 5}{0.247/\sqrt{9}} = -2.562$$

    - the 99% confidence limit of the t statistics with 8 d.o.f. are: $t_{8,lim} = [-3.355, 3.355]$
    - we fail to reject the null hypothesis (i.e., H_0 is accepted) because the observed value falls within the acceptance region

$$t_o = -2.562 \in [-3.355, 3.355]$$

Example of hypothesis testing on Student t: water contamination (2/2)

- The observed value t_o falls within the acceptance region $|t_o| < t_{8,lim}$, so the null hypothesis H_0 cannot be rejected with confidence $\alpha = 0.01$



Example of hypothesis testing on χ^2: enzymatic activity

- Example:
  - you work for a biopharmaceutical multinational
  - your company wants to maintain the variability of the enzymatic activity for a drug manufacturing process lower than $0.0015u^2$ with a significance $\alpha = 0.05$
    - the data are stored in the file: **`enzymatic_activity.xlsx`**

- Solution:
  - the hypothesis to be tested is:

$$\begin{cases} \mathrm{H}_0\colon \sigma^2 = 0.0015 \\ \mathrm{H}_1\colon \sigma^2 < 0.0015 \end{cases}$$

1-sided test because we want to exclude only that $\sigma^2 < 0.0015$

  - a χ^2-test should be performed:
    - the observed value of χ^2 is:

$$\chi_o^2 = \frac{(N-1)S^2}{\sigma^2} = \frac{(100-1)0.000985}{0.0015} = 65.01$$

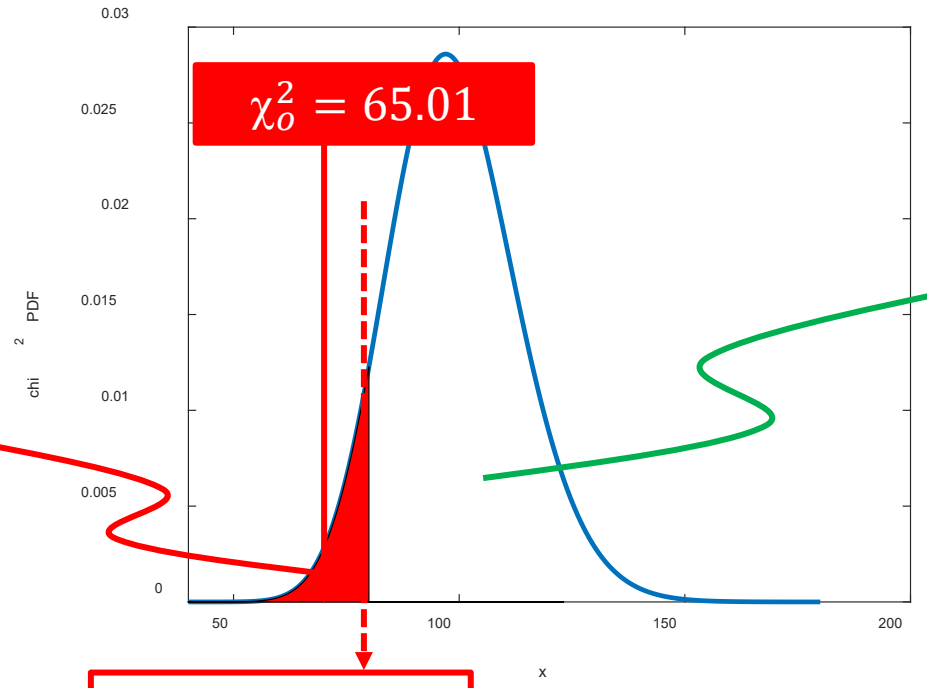
    - the 5% confidence limit of the χ^2 statistics with 99 d.o.f. (referred to the left tail only) is: $\chi^2_{99,lim} = 77.05$
    - the null hypothesis is rejected because the observed value falls in the acceptance region

$$\chi_o^2 = 65.01 < 77.05$$

Example of hypothesis testing on χ^2: enzymatic activity

- The null hypothesis is rejected because the observed value falls out of the acceptance region, whose limite at $\alpha = 0.05$ is $\chi^2_{99,lim} = 77.05$:

$$\chi^2_o = 65.01 \notin [77.05, +\infty[$$



rejection region in the left tail only, because a one-sided test is performed

acceptance region

Example of hypothesis testing on Fisher F: pasteurization

- Example:
  - you work in a food company performing food pasteurization
  - you want a mathematical demonstration that the 2 ovens that are utilize to pasteurize food are equivalent with a significance $\alpha = 0.05$
    - the data are stored in the file: **`pasteurization.xlsx`**

- Solution:
  - the hypothesis to be tested is:

$$\mathrm{H}_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$\mathrm{H}_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

  - a F-test should be performed:
    - the observed value of F is:

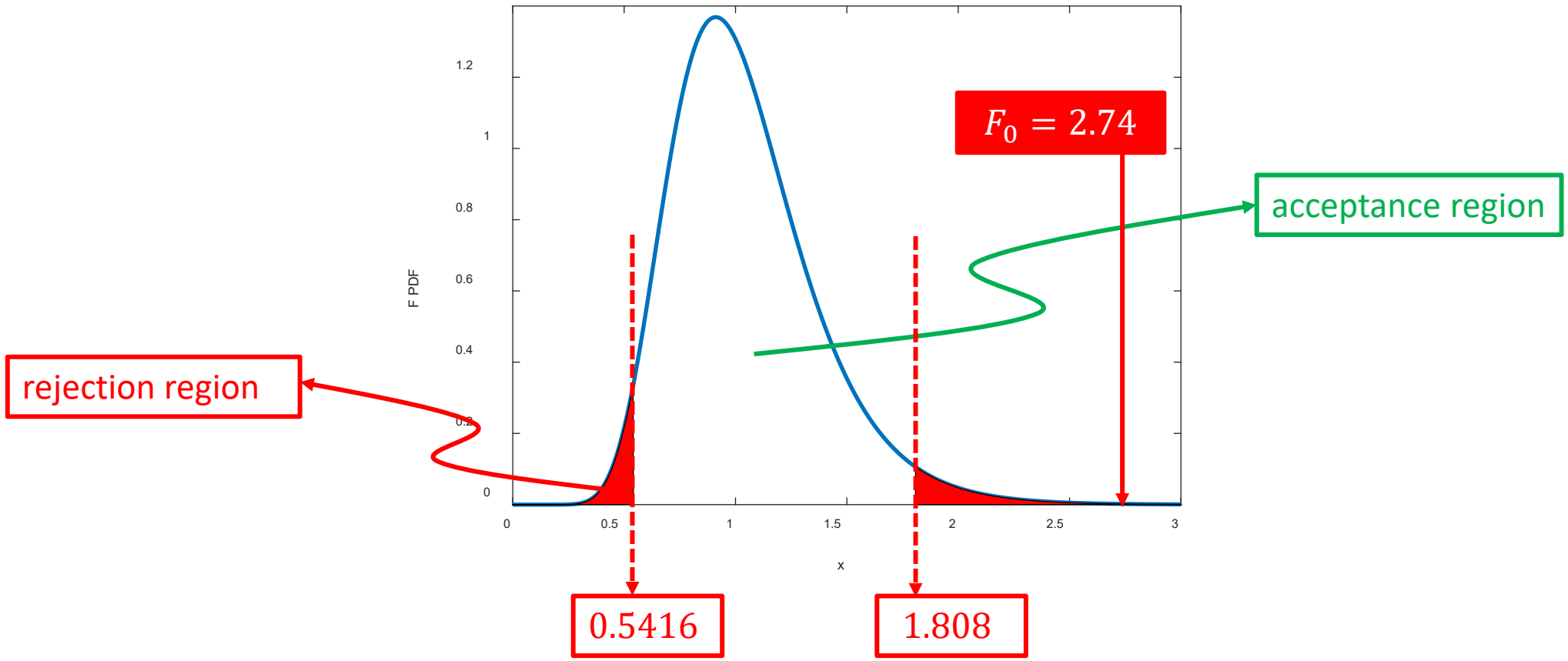
$$F_0 = \frac{\sigma_1^2}{\sigma_2^2} = 2.74$$

    - the 95% confidence limits of the F statistics are: $F_{lim} = [0.5416, 1.808]$
    - the null hypothesis is rejected because the observed value does not fall within the acceptance region

$$F_0 = 2.74 \notin [0.5416, 1.808]$$

Example of hypothesis testing on Fisher F: pasteurization

- The null hypothesis is rejected because the observed value does not fall within the acceptance region $F_0 = 2.74 \notin [0.5416, 1.808]$



What we learned in the first lessons

- We learned a lot about **low complexity** problems on a single variable **variability** (whatever its **nature** is):
  - concepts:
    - random variables
    - probability
    - probability and cumulative density function
    - inverse density function
    - inference: mean, variance, percentiles, moments, etc…
    - central limit theorem
  - distributions:
    - **normal and standard normal distribution**
    - χ^2 **distribution**
    - **Students'** t **distribution**
    - **Fisher's** F **distribution**
  - how to calculate:
    - descriptions and verifications on a distribution
    - the probability that a determined event occur
    - the event associated with a predetermined chance of occurring

generated with ChatGPT

Applications

With Minitab®

1-sample T test

Example of 1-sample T test

- **Problem**:
  - same problem as in the previous case

Hypothesis testing formulation

- Hypothesis test:

$$H_0: \mu = 5$$
$$H_1: \mu \neq 5$$

- How could we solve this problem with Minitab®?
  1. calculate the observed t
  2. observe if it falls in the rejection zone
     - perform a 1-sample t test

Solution: strategy #1

- Calculate the observed t:

$$t_{\text{obs}} = \frac{\bar{x} - \mu}{s/\sqrt{N}} = \frac{4.789 - 5}{0.247/\sqrt{9}} = -2.562$$

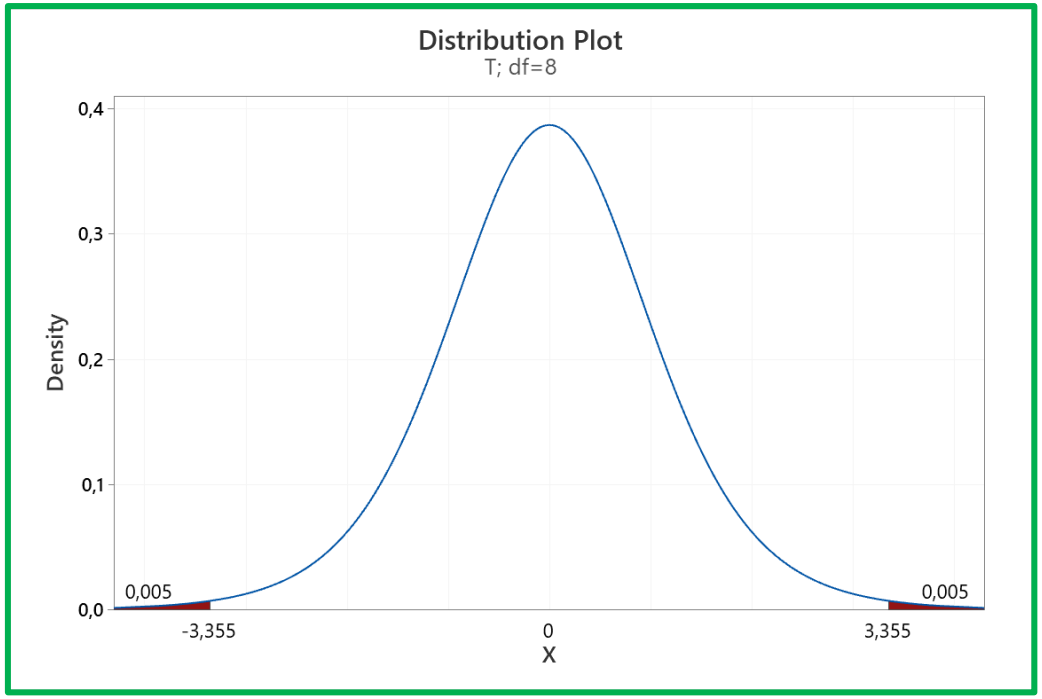
- Calculate the $\alpha = 0.01$ confidence limit in Minitab®:
  - **Graph**
  - **Probability distribution plot**
  - click **View Probability**
  - OK
  - select **Distribution: t**
  - select **Degrees of freedom: 8**
  - click **Shaded area**
  - select **Probability**
  - **Both tails**
  - impose the probability value at **0.01**

Results

- Compare the observed t with the limiting values:

$$t_{\text{obs}} = 2.562 < 3.355$$

- **We fail to reject (i.e., we accept) the null hypothesis**



And without the software…

- Search the value in the textbook tables

ν \ α	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.727	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.019	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781

Solution: (lazy) strategy #2

- **Stat**
- **Basic statistics**
- **1-sample t**
- Select the dataset
- Select **One or more samples…**
- Click **Perform hypothesis test**
- Impute **Hypothesized mean: 5**
- Select **Options**
- Select **Confidence level: 0.99**
- OK
- Select **Alternative hypotesis: Mean ≠ hypothesized mean**
- Select **Graph**: all the plots
- Click OK twice

Results

- The observed t is within the limits
  - it is in the acceptance region
- **We fail to reject the null hypotesis:**
  - **the sample mean coincide, from the sstatistical point of view, with the expected one!**

Descriptive Statistics

N	Mean	StDev	SE Mean	99% CI for μ
9	4.7889	0.2472	0.0824	(4.5124; 5.0654)

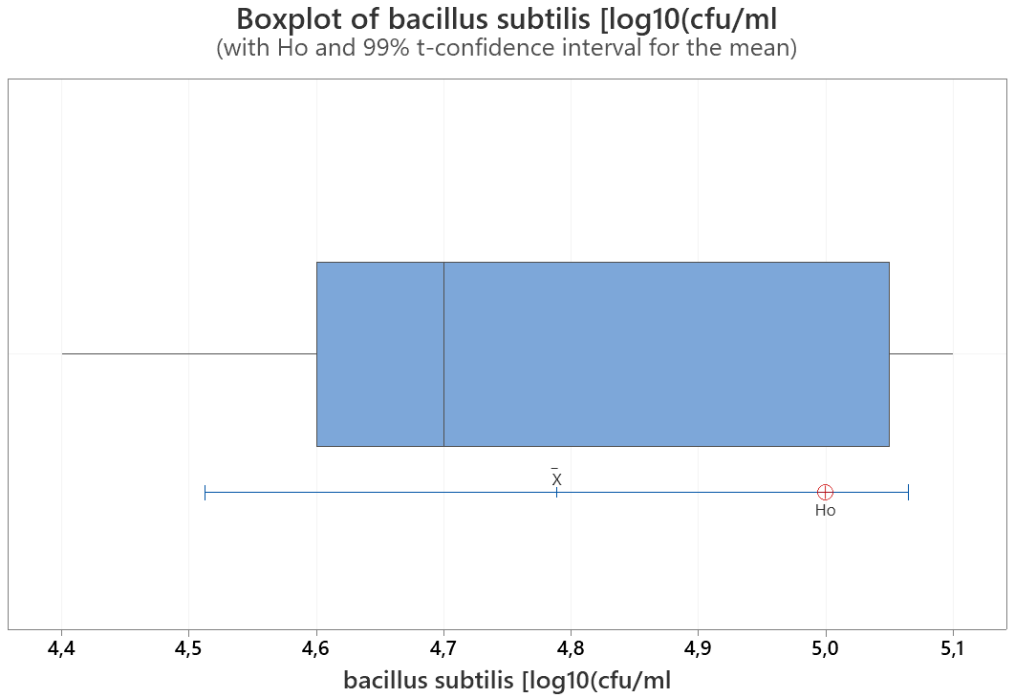
μ: population mean of bacteria [log10(cfu/ml)]

Test

Null hypothesis $H_0: \mu = 5$

Alternative hypothesis $H_1: \mu \neq 5$

T-Value	P-Value
-2,56	0,034



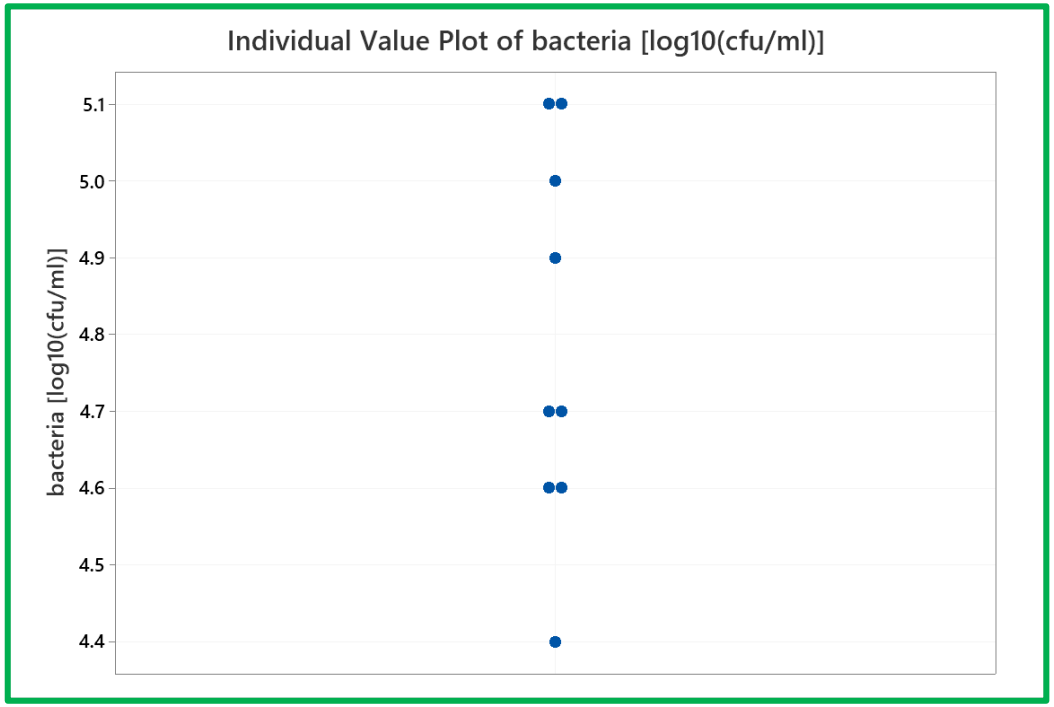
1-sample Z test

Example of 1-sample Z test

- **Problem**: determine if the value of bacteria [$\log_{10}$(cfu/ml)] contamination in water is 5 with a confidence of $\alpha = 0.01$, as prescribed by law
- Available data:
  - $N = 9$ observations are taken in one sample
    - assume a Gaussian distribution with $\sigma = 0.2$
    - dataset:
      - `water_contamination.xlsx`

Dataset visualization

- Preliminary data visualization:
  - **`Graph`**
  - **`Individual value plot`**
  - click OK
  - select the variable
  - click OK



Hypothesis testing formulation

- Hypothesis test:

$$H_0: \mu = 5$$
$$H_1: \mu \neq 5$$

- How could we solve this problem with Minitab®?
  1. calculate the observed z
  2. observe if it falls in the rejection zone
     - perform a 1-sample Z test

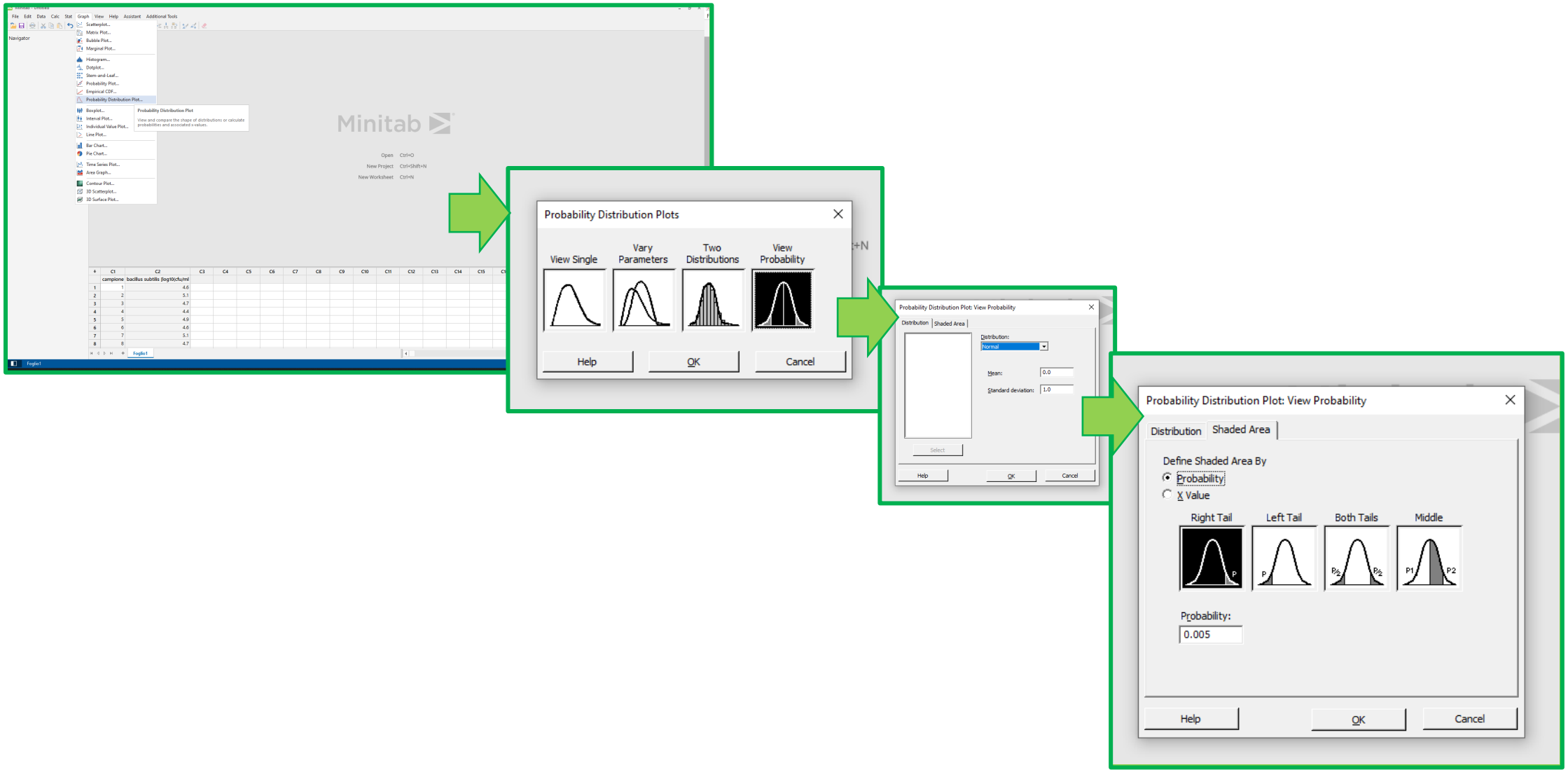
Solution: strategy #1

- Pen and paper:
  - calculate the observed value for z:

$$z_{\text{obs}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} = \frac{4.789 - 5}{0.2/\sqrt{9}} = -3.167$$

- Calculate the confidence limit of z for $\alpha = 0.01$ with Minitab®
  - **Graph**
  - **Probability distribution plot**
  - **View Probability**
  - OK
  - select **Distribution: normal**
    - mean **0**
    - standard deviation **1**
  - click **Shaded area**
  - select **Probability**
  - **Both tails**
  - digit the value of probability **0.01**

Implementation in Minitab®

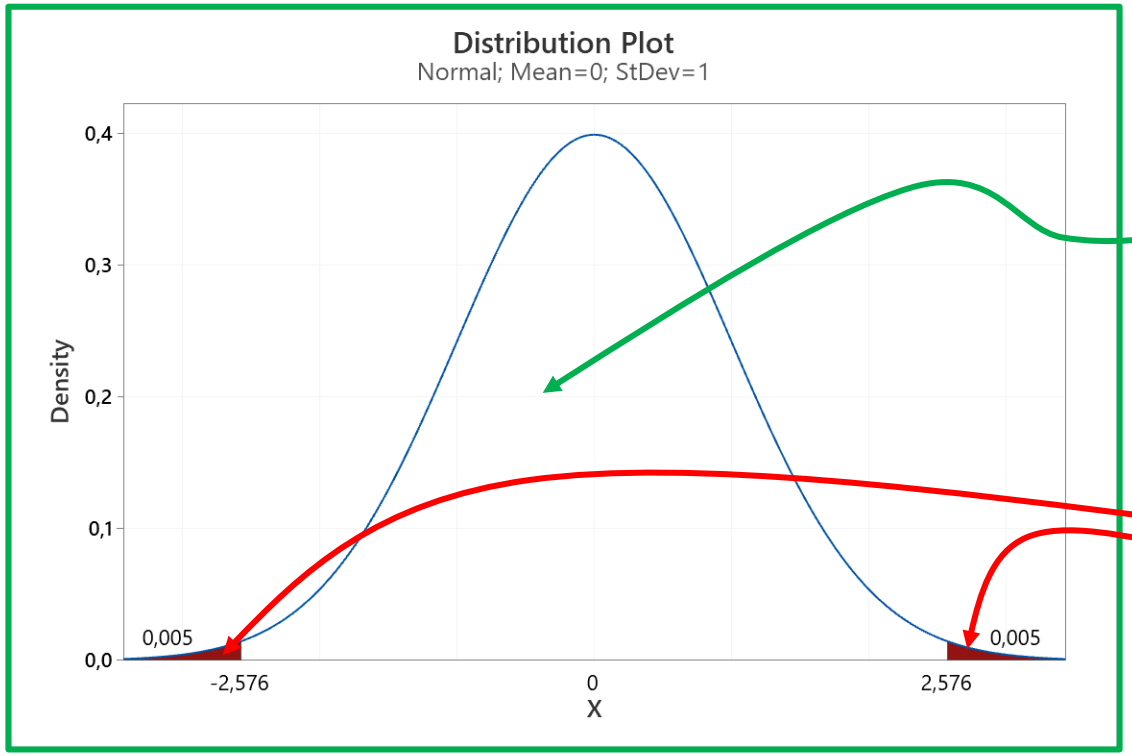


Results

- Compare the observed z value and the limits:

$$Z_{\text{obs}} = -3.167 < -2.576$$

- **The null hypothesis is rejected because the observed value is in the rejection region**

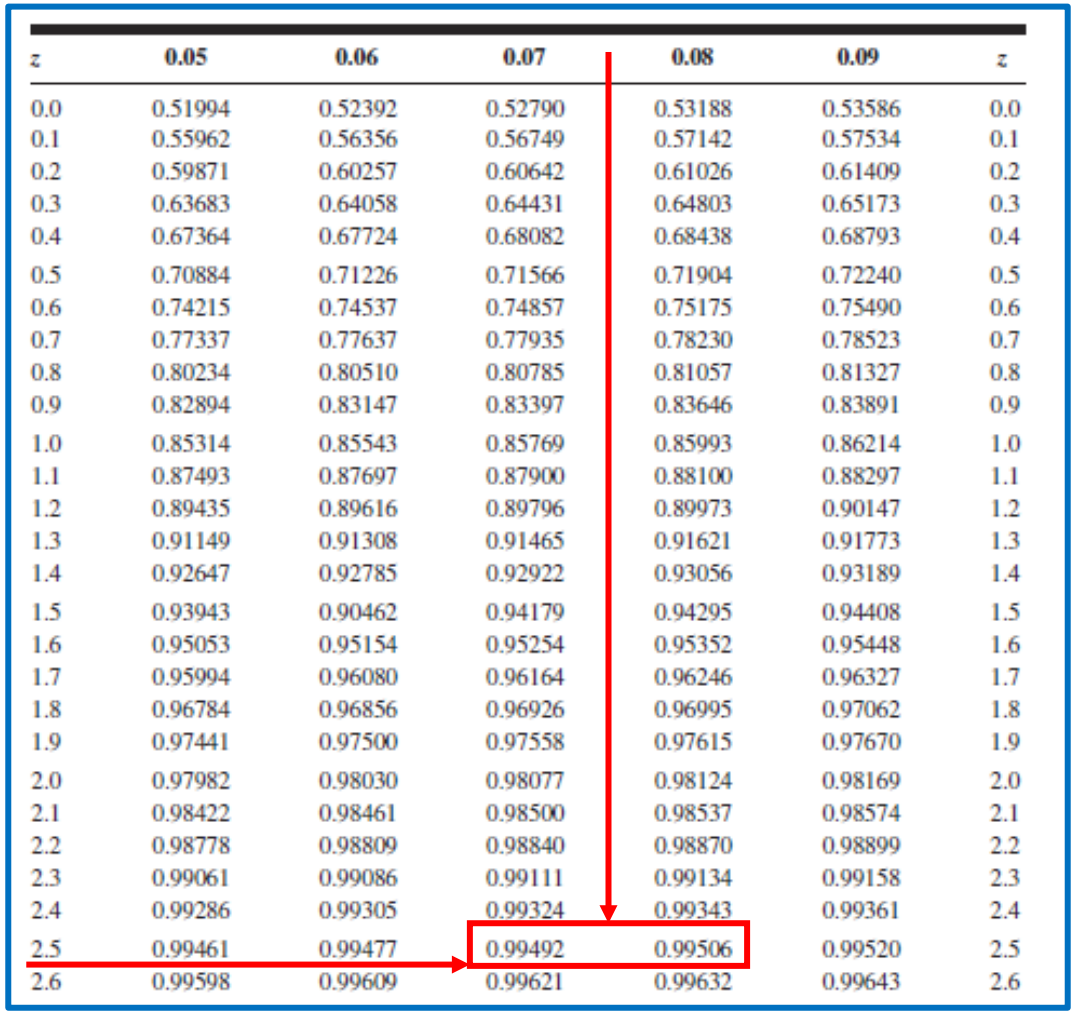


white acceptance region

red rejection regions

… and without Minitab®?

- Search the value in the textbook tables

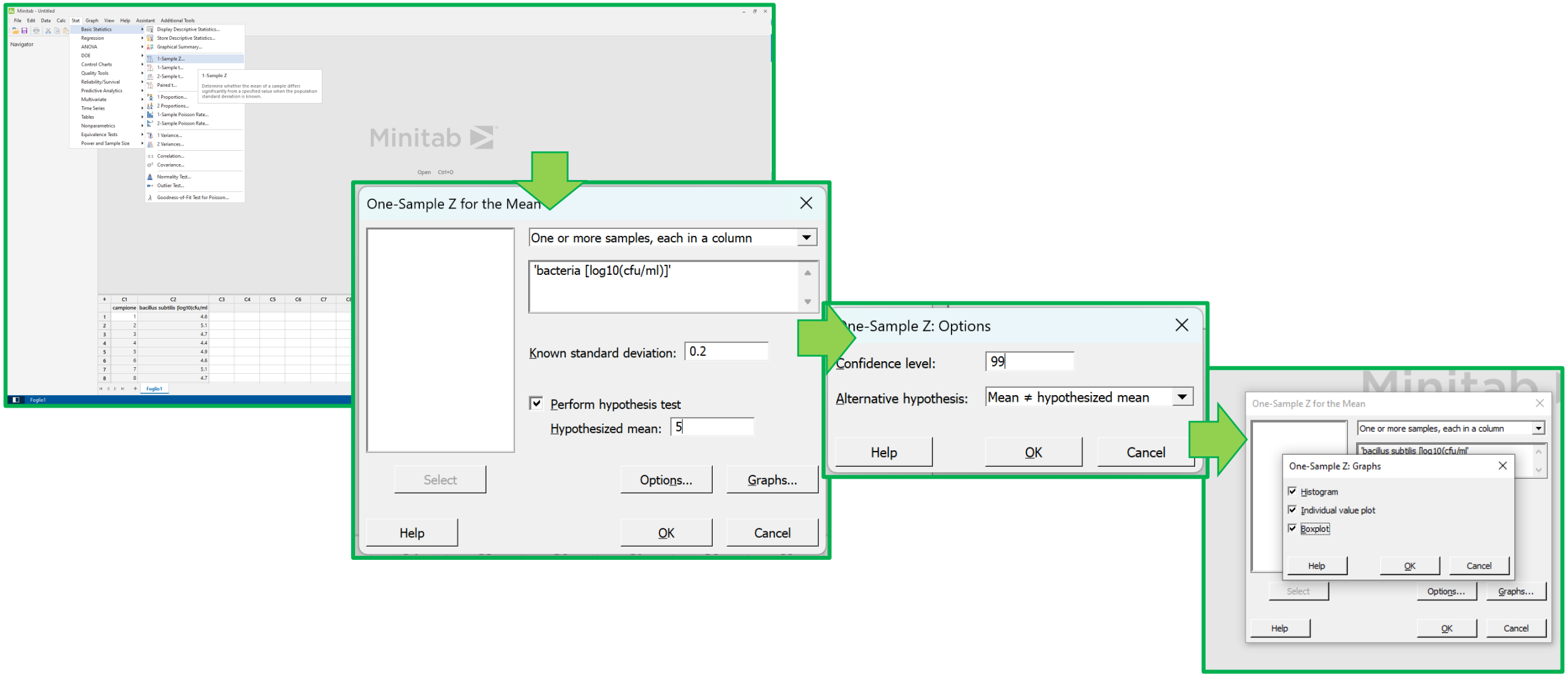


z	0.05	0.06	0.07	0.08	0.09	z
0.0	0.51994	0.52392	0.52790	0.53188	0.53586	0.0
0.1	0.55962	0.56356	0.56749	0.57142	0.57534	0.1
0.2	0.59871	0.60257	0.60642	0.61026	0.61409	0.2
0.3	0.63683	0.64058	0.64431	0.64803	0.65173	0.3
0.4	0.67364	0.67724	0.68082	0.68438	0.68793	0.4
0.5	0.70884	0.71226	0.71566	0.71904	0.72240	0.5
0.6	0.74215	0.74537	0.74857	0.75175	0.75490	0.6
0.7	0.77337	0.77637	0.77935	0.78230	0.78523	0.7
0.8	0.80234	0.80510	0.80785	0.81057	0.81327	0.8
0.9	0.82894	0.83147	0.83397	0.83646	0.83891	0.9
1.0	0.85314	0.85543	0.85769	0.85993	0.86214	1.0
1.1	0.87493	0.87697	0.87900	0.88100	0.88297	1.1
1.2	0.89435	0.89616	0.89796	0.89973	0.90147	1.2
1.3	0.91149	0.91308	0.91465	0.91621	0.91773	1.3
1.4	0.92647	0.92785	0.92922	0.93056	0.93189	1.4
1.5	0.93943	0.90462	0.94179	0.94295	0.94408	1.5
1.6	0.95053	0.95154	0.95254	0.95352	0.95448	1.6
1.7	0.95994	0.96080	0.96164	0.96246	0.96327	1.7
1.8	0.96784	0.96856	0.96926	0.96995	0.97062	1.8
1.9	0.97441	0.97500	0.97558	0.97615	0.97670	1.9
2.0	0.97982	0.98030	0.98077	0.98124	0.98169	2.0
2.1	0.98422	0.98461	0.98500	0.98537	0.98574	2.1
2.2	0.98778	0.98809	0.98840	0.98870	0.98899	2.2
2.3	0.99061	0.99086	0.99111	0.99134	0.99158	2.3
2.4	0.99286	0.99305	0.99324	0.99343	0.99361	2.4
2.5	0.99461	0.99477	0.99492	0.99506	0.99520	2.5
2.6	0.99598	0.99609	0.99621	0.99632	0.99643	2.6

Solution: strategy #2

- If you are lazy, you can perform the test in a fully automatic fashion:
  - **Stat**
  - **Basic statistics**
  - **1-sample Z**
  - select the dataset
  - select **One or more samples…**
  - write **Known standard deviation: 0.2**
  - click **Perform hypothesis test**
  - impute **Hypothesized mean: 5**
  - select **Options**
  - select **Confidence level: 99**
  - select **Alternative hypothesis: Mean ≠ hypothesized mean**
  - click OK
  - select **Graph**: all the plots
  - click OK twice

Implementazione in Minitab®



Results

- The observed z is out of the acceptance region
- The p-value is very low
  - the observed value is far from the limits of the acceptance region
- **The null hypothesis must be rejected:**
  - **the sample mean does not coincide, from the statistical point of view, with the expected value $\mu = 5$**

Descriptive Statistics

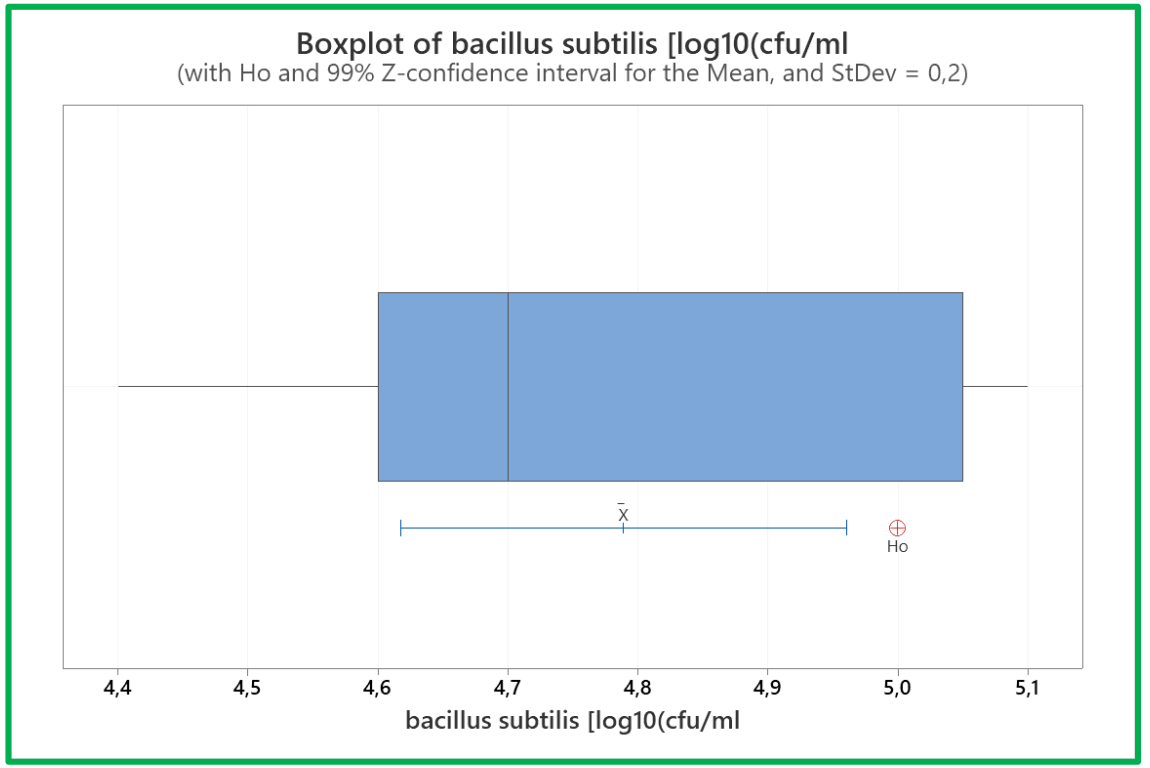
N	Mean	StDev	SE Mean	99% CI for μ
9	4.7889	0.2472	0.0667	(4.6172; 4.9606)

μ: population mean of bacteria [log10(cfu/ml)]
Known standard deviation = 0.2

Test

Null hypothesis H_0: μ = 5
Alternative hypothesis H_1: μ ≠ 5

Z-Value	P-Value
-3,17	0,002



1-variance test

Example of 1 variance test

- **Problem**:
  - you work in the field of biopharm
  - your company wants to maintain the variability of the enzymatic activity for a drug manufacturing process lower than $0.0015u^2$ with a confidence of $\alpha = 0.05$

- **Question**:
  - verify that the variance is at the expected value with a confidence level of $\alpha = 0.05$
  - dataset: `enzymatic_activity.xlsx`

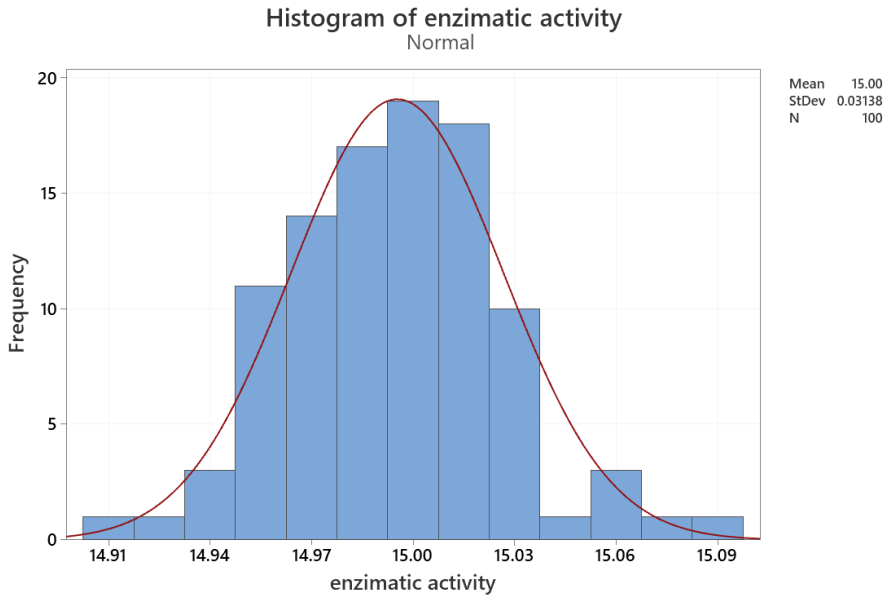
Exploratory analysis and data visualization

- Preliminary steps:
  - **descriptive statistics analysis** of the dataset
  - data visualization through a **histogram**
- In Minitab®:
- descriptive statistics
  - `Stat`
  - `Basic statistics`
  - `Display descriptive statistics`
  - select `Statistics`: select all
  - click OK
  - select `Graphs`: select all
  - click OK twice

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Variance	CoefVar	Minimum	Q1
enzimatic activity	100	0	14.995	0.00314	0.0314	0.00099	0.21	14.913	14.971

Variable	Median	Q3	Maximum	IQR	Skewness	Kurtosis
enzimatic activity	14.995	15.014	15.089	0.0427	0.27	0.41



Hypothesis testing formulation

- The hypothesis test is:

$$H_0: \sigma^2 = 0.0015$$

$$H_1: \sigma^2 < 0.0015$$

- How could we solve this problem with Minitab®?
  - perform a 1-variance test

Solution

- **Stat**
- **Basic statistics**
- **1 variance**
- Select the dataset
- Select **One or more samples…**
- Select **Perform hypothesis test**
- Select **Hypothesized variance** and impute **0.0015**
- Select **Options**
- Select **Confidence level: 0.95**
- Select **Alternative hypothesis: Variance < hypothesized variance**
- Click OK twice

Results

- The computed χ^2
- Bonnet test and χ^2 test:

$$\chi^2 = \frac{(N-1)S^2}{\sigma^2} = \frac{(100-1)0.000985}{0.0015} = 65.01$$

- The respective p-value
- **The null hypothesis is rejected, and the enzymatic activity is statistically lower than 0.0015**

Descriptive Statistics

N	StDev	Variance	95% Upper Bound for σ using Bonett	95% Upper Bound for σ using Chi-Square
100	0,0314	0,000985	0,0360	0,0356

Test

Null hypothesis $H_0: \sigma^2 = 0,0015$

Alternative hypothesis $H_1: \sigma^2 < 0,0015$

Method	Test Statistic	DF	P-Value
Bonett	—	—	0,007
Chi-Square	65,01	99	0,003

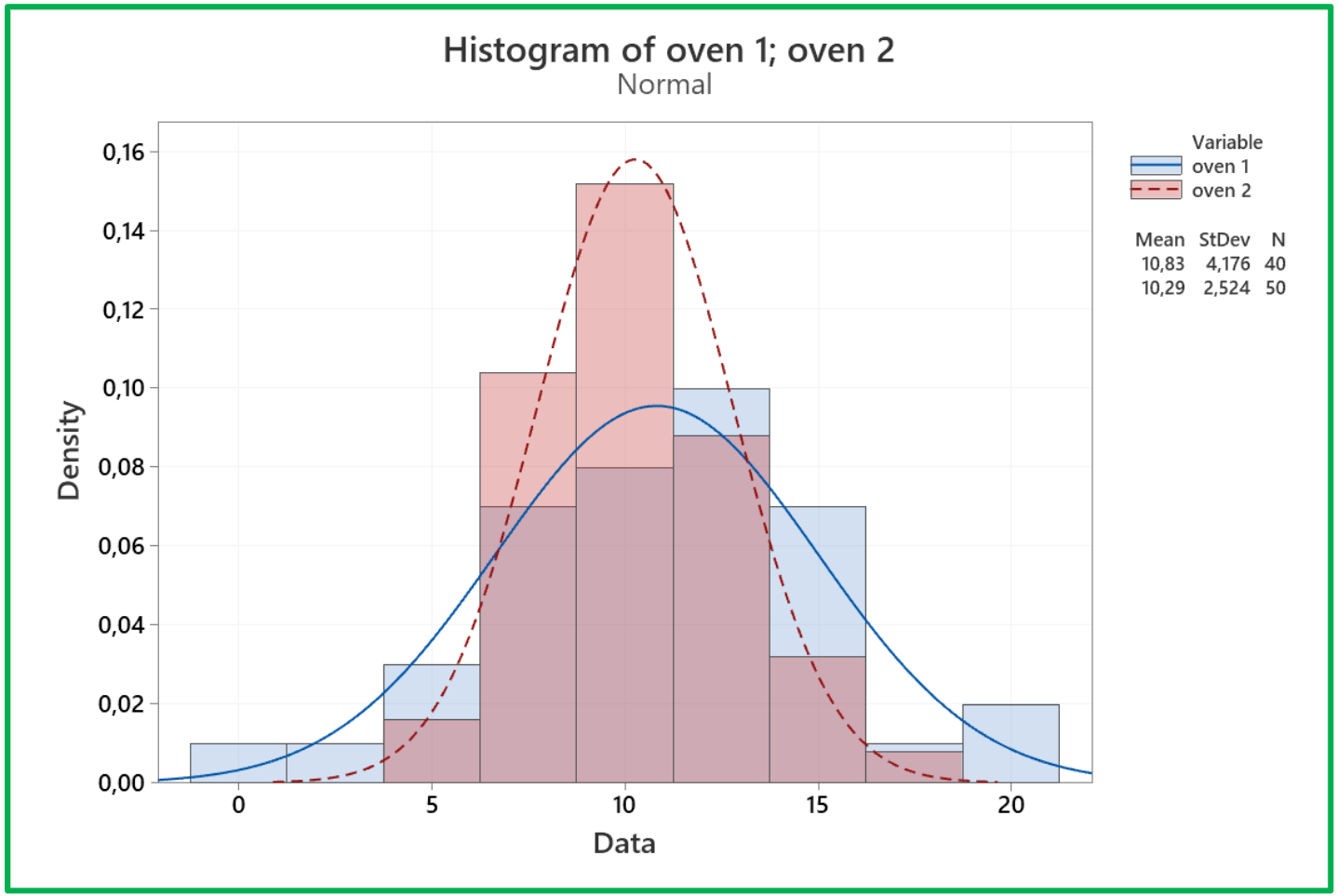
2-variance test

Example of 2 variance test

- **Problem**: PureTomato is a company working on the field of food pasteurization. They use 2 different ovens for the pasteurization of tomato sauce with heat.
- **Question**: do the ovens have the same performance in terms of variability of the bacterial contamination
  - dataset: `pasteurization.xlsx`

Data visualization

- **Graph**
- **Histogram**
- **With fit and groups**
- Select the ovens outcomes
- Click OK



Hypothesis testing formulation

- We want to understand if the 2 ovens pasteurize the tomato sauce in different manner:

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

- How could we solve this problem with Minitab®?
  1. calculate the observed F value
  2. observe if it falls in the rejection zone
     - perform a 2-variance test

Solution: strategy #1

- Calculate the observed F:

$$F_{\text{obs}} = \frac{\sigma_1^2}{\sigma_2^2} = 2.74$$

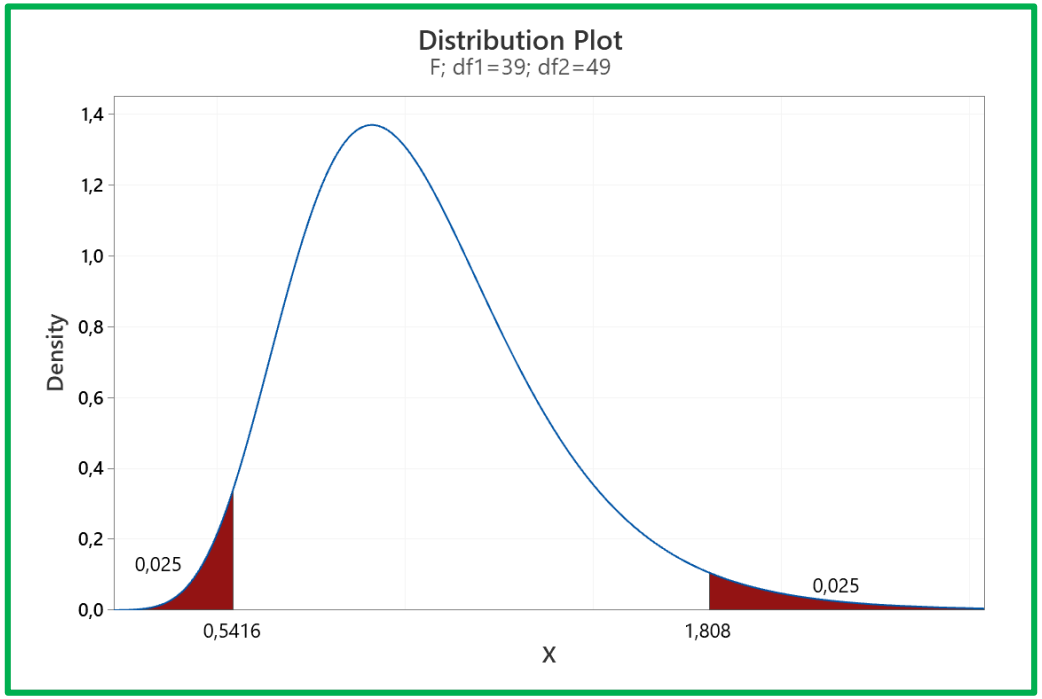
- Calculate the confidence limits of F with $\alpha = 0.05$ in Minitab®:
  - **Graph**
  - **Probability distribution plot**
  - click **View Probability**
  - click OK
  - select **Distribution: F**
    - number of d.o.f.:
      - **Numerator df: 39**
      - **Denominator df: 49**
  - click **Shaded area**
  - select **Probability**
  - **Both tails**
  - input the probability value: **0.05**
  - click OK

Results

- Comparing the observed statistics with the confidence limits:

$$F_{\text{obs}} = 2.74 > 1.808$$

- **The null hypothesis is rejected:**
  - **the 2 ovens perform different pasteurization**



Solution: strategy #2

- **Stat**
- **Basic statistics**
- **2 Variance**
- Select the dataset
- Select **Each sample is in its own column**
- Select the ovens data
- Select **Options**
- Select **Ratio: (sample 1 variance)/(sample 2 variance)**
- Select **Confidence level: 95**
- Select **Hypothesized ratio: 1**
- Select **Alternative hypothesis: Ratio ≠ hypothesized ratio**
- Select **Use test and confidence interval…**
- Click OK twice

Results

- The observed F is out of the confidence limits
- The p-value is low
- **The null hypothesis is rejected**

Descriptive Statistics

Variable	N	StDev	Variance	95% CI for σ
oven 1	40	4,176	17,443	(3,421; 5,363)
oven 2	50	2,524	6,373	(2,109; 3,146)

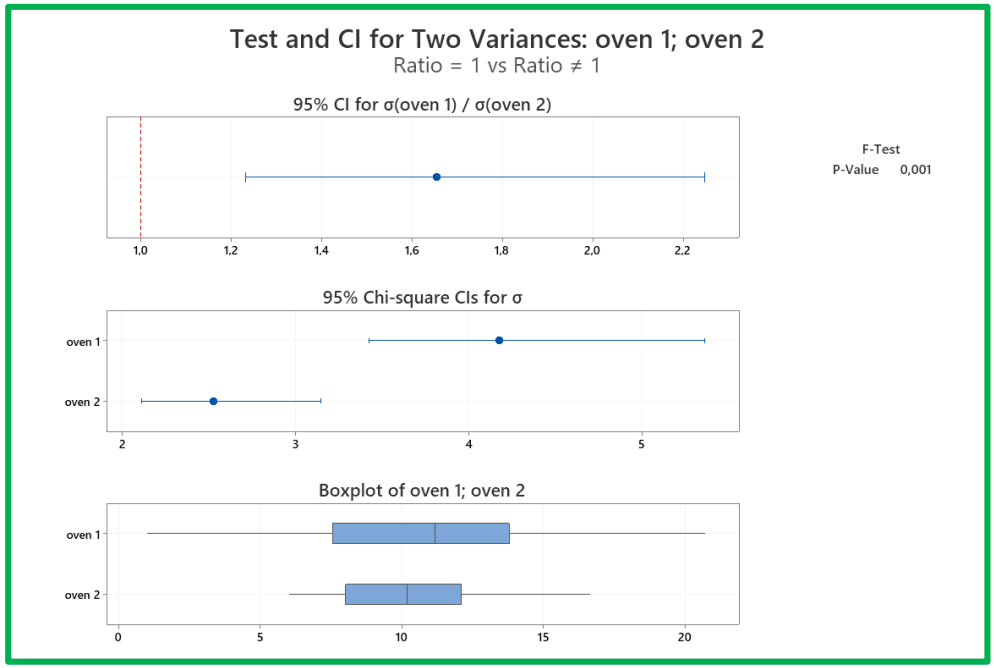
Ratio of Standard Deviations

Estimated Ratio	95% CI for Ratio using F
1.65444	(1.230; 2.248)

Test

Null hypothesis	$H_0: \sigma_1 / \sigma_2 = 1$
Alternative hypothesis	$H_1: \sigma_1 / \sigma_2 \neq 1$
Significance level	$\alpha = 0.05$

Method	Test Statistic	DF1	DF2	P-Value
F	2.74	39	49	0.001



… per sempre a fianco a me!